Adv in Health Sci Educ (2010) 15:647–658
DOI 10.1007/s10459-010-9225-8

Using signal detection theory to model changes in serial learning of radiological image interpretation

Kathy Boutis · Martin Pecaric · Brian Seeto · Martin Pusic

Received: 13 October 2009 / Accepted: 1 February 2010 / Published online: 26 February 2010

Abstract Signal detection theory (SDT) parameters can describe a learner's ability to discriminate (d') normal from abnormal and the learner's criterion (λ) to under or overcall abnormalities. To examine the serial changes in SDT parameters with serial exposure to radiological cases. 46 participants were recruited for this study: 20 medical students (MED), 6 residents (RES), 12 fellows (FEL), 5 staff pediatric emergency physicians (PEM), and 3 staff radiologists (RAD). Each participant was presented with 234 randomly assigned ankle radiographs using a web-based application. Participants were given a clinical scenario and considered 3 views of the ankle. They classified each case as normal or abnormal. For abnormal cases, they specified the location of the abnormality. Immediate feedback included highlighting on the images and the official radiologist's report. The low experience group (MED, RES, FEL) showed steady improvement in discrimination ability with each case, while the high experience group (PEM, RAD) had higher and stable discrimination ability throughout the exercise. There was also a difference in the way the high and low experience groups balanced sensitivity and specificity (λ) with the low experience group tending to make more errors calling positive radiographs negative. This tendency was progressively less evident with each increase in expertise level. SDT metrics provide valuable insight on changes associated with learning radiograph interpretation, and may be used to design more effective instructional strategies for a given learner group.

K. Boutis (✉)
Department of Pediatrics, The Hospital for Sick Children, University of Toronto,
555 University Avenue, Toronto, ON M5G 1X8, Canada
e-mail: boutis@pol.net

M. Pecaric
Contrail Consulting Services, Toronto, ON, Canada

B. Seeto
School of Medicine, Queen's University, Kingston, ON, Canada

M. Pusic
Department of Pediatrics, Morgan Stanley Children's Hospital, Columbia University,
New York, NY, USA

Keywords Image interpretation · Signal detection theory

Introduction

Real-life medical decisions usually have to be made based on the weight of the clinical evidence with some inherent uncertainty. Signal detection theory (SDT) allows the quantitative description of an observer's decision making by considering two separate aspects. The first is called discrimination and measures how well the observer is able to make correct judgements and avoid incorrect ones. The second is called the criterion, which is a measure of the bias with which the observer favours one option over another, independent of the ability to discriminate (McNicol 2004; Wickens 2002).

In the context of radiograph interpretation, discrimination (discrimination parameter, d') assesses one's ability to discern normal from abnormal radiographs. In Fig. 1, the novice diagnostician has difficulty telling the difference between normal and abnormal films, and therefore d' is small. Meanwhile, an expert can, to a much greater extent, separate the images in terms of their degree of abnormality (d' is large). The criterion (criterion parameter, λ) measures the inclination to classify a given case as normal or abnormal, especially for borderline cases. Clinicians with the same ability to discriminate, viewing the same radiographs, may have a different propensity to call the radiograph normal. There is no single correct value of the criterion parameter that someone who is interpreting a radiograph should adopt. The criterion that one performs with depends on the goal (s)he has in mind, and this likely varies with a given situation (McNicol 2004). A higher (strict) criterion trades high specificity for relatively lower sensitivity, while a lower criterion does the opposite (Fig. 2). For example, the radiograph reader may wish to maximize the number of cases (s)he calls abnormal in order to avoid missing significant pathology. Alternatively, one may be biased towards assuming that most radiographs will be normal while reviewing radiographs, thereby resulting in a more specific λ. The most efficient criterion parameter neither under or overcalls radiographs as normal or abnormal.

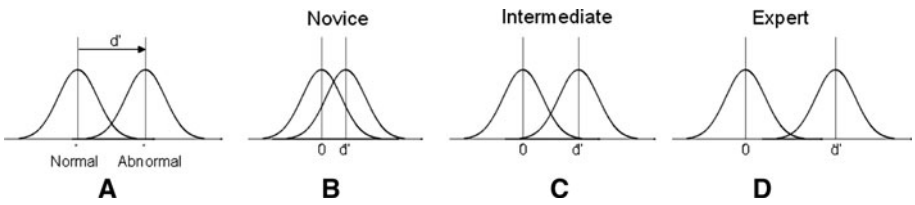


Fig. 1 Signal detection parameter discrimination—reported as dPrime (d'). The x-axis represents a latent factor to be detected (in the case of radiograph interpretation, the degree of abnormality). The y-axis is the number of examples. **a** A representation of the equal variance signal detection model. For radiograph interpretation, the left hand curve represents normal cases while the right hand curve cases with abnormalities. The subject's ability to discriminate is measured as the distance between their means of these distributions, quantified as d'. **b** The novice radiograph interpreter has difficulty distinguishing abnormal from normal radiographs. Therefore, the separation of the means of the two distributions is small and hence the subject's d' value is small. **c** The intermediate is better able to distinguish normal from abnormal and hence their d' is larger. However, there is still considerable overlap, cases where they are unable to reliably classify the radiograph. **d** The expert situation: d' is large with relatively little overlap (few cases where the expert cannot distinguish normal from abnormal)

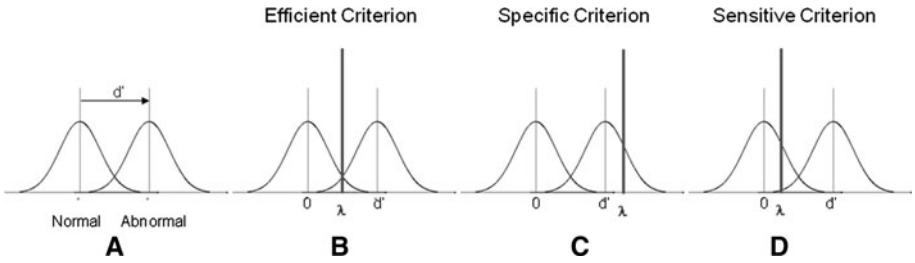


Fig. 2 Signal detection parameter criterion—reported as lambda (λ). **a** As stated in Fig. 1. **b** The criterion, λ, is the subject's bias to call a radiograph for which they are uncertain. In **b**, the subject has set a criterion that maximizes their accuracy, for a given ability to discriminate. **c** A highly specific criterion favours being correct for Normals over being correct for Abnormals. In this case, $\lambda > d'$. Accuracy suffers due to false negatives. **d** In this panel, we show a highly sensitive criterion where the subject favours being correct for abnormals. Numerically, λ is generally less than $d'/2$

In medicine, SDT has been mostly applied to the clinical aspects of diagnostic radiology (Clarkson 2007; Doubilet 1988; Metz 1986; Swensson 1996; Abdi 2009; Norman et al. 1992). SDT is based on the familiar 2 × 2 table used to calculate sensitivity and specificity, and in radiology this same fundamental concept has been represented using Receiver Operating Characteristic (ROC) curves. The equivalent to the discrimination parameter is the "area under the curve," while the point on the ROC curve for a given individual is the homologue of the criterion parameter (Obuchowski 2003).

While these concepts have been used for assessment in radiology education (Clarkson 2007), what has been uncommon is their use in a *serial* fashion where estimates of discrimination and criterion are calculated at regular intervals as a trainee develops their ability to interpret radiographs. Collecting an estimate of a learners' ability to classify radiographs as they learn results in a learning curve (Fig. 3) (Ericsson 2006; Ericsson et al. 1993; Hatala et al. 2003; Ramsay et al. 2001; Nodine et al. 1999), and allows us to draw conclusions about the nature of their individual learning at a fine level of granularity.

The main objective of this study was to examine how a learner's discrimination and criterion change with serial exposure to radiological cases in order to better understand the development of competency in radiograph interpretation.

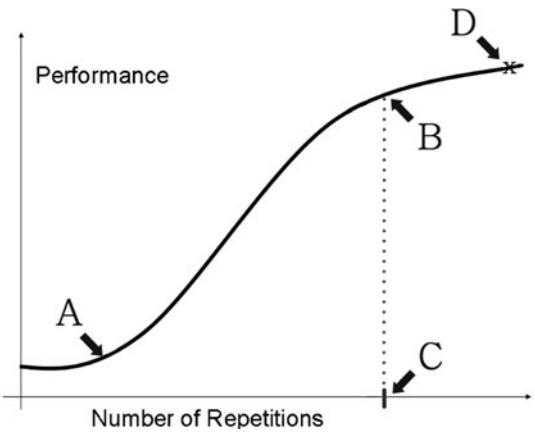


Fig. 3 The learning curve—the *x*-axis represents the number of repetitions or period of time spent learning while the *y*-axis is some index of performance, typically accuracy rate. *Point A* number of repetitions at which learning begins (participant is oriented and accustomed to the educational intervention. *Slope AB* initially rapid rate of learning; *Point C* inflection point at which learning becomes more effortful; *Point D* total number of repetitions required to achieve a given level of competency

Methods

Participant recruitment

We recruited a convenience sample of participants for this study. The following individuals were contacted via electronic mail with an opportunity to participate: final year medical students from three medical schools (University of Toronto, Queen's University, Columbia University) rotating through the emergency department ($n = 56$), senior pediatric and emergency residents ($n = 10$), emergency pediatric fellows ($n = 30$), staff emergency physicians ($n = 20$), and staff pediatric radiologists ($n = 10$) from two Children's Hospitals (The Hospital for Sick Children and Morgan Stanley Children's Hospital). The medical students (MED), residents (RES), and fellows (FEL) were considered the "low experience group" while staff emergency physicians (PEM) and radiologists (RAD) were considered the "high experience group." This study was approved by the research ethics boards at the participating institutions.

Radiograph selection and diagnostic classification

Ankle films were chosen as the type of radiographs for this image bank because they represent one of the most common pediatric injuries and available evidence demonstrates that physicians lack skills to manage musculoskeletal injuries appropriately (Chung et al. 2004; Dowling et al. 2005; Minnes et al. 2005; Reeder et al. 2004; Ryan et al. 2004; Taras 1990; Trainor and Krug 2000). A clinician must make a dichotomous decision based on the clinical information. That is, based on the radiograph interpretation, the clinician must either (a) declare the radiograph free of fracture and discharge the patient with only supportive measures or (b) diagnose a fracture and appropriately manage the patient with splinting and arrangements for further care. On average, a pediatric emergency department sees one ankle injury requiring radiographs per day (Boutis et al. 2001). Given the average pediatric emergency physician in Canada works approximately 24 clinical hours per week, 200 radiographs replicates a 4 year experience of a practicing emergency pediatrician. Therefore, our target was to collect an image bank of approximately 200 radiographs, which is also in keeping with the minimum number of trials required for SDT research (McNicol 2004).

We collected the radiographs by purposively sampling from our clinical setting. Over a 1 year period (September 2005–August 2006), we assembled a database of 378 consecutive ankle radiographs (AXR) taken in a pediatric emergency department (PED) for the purpose of excluding a fracture. From these, we selected 234 AXR using the following process: in the first step, we excluded all films that had orthopedic hardware (12), markers like casting material that would suggest a diagnosis (71), very poor quality films such that radiograph findings were obscured (7), and radiographs where the diagnosis changed in follow up films (26). This resulted in a pool of 261 available radiographs from which two pediatric emergency medicine (PEM) staff physicians (KB, MP) then selected 234 radiographs which provided the following content: a case-mix frequency of abnormal/normal/normal variant radiographs consistent with that seen in actual clinical practice (Boutis et al. 2001) and emphasized necessary educational content known for emergency management of pediatric ankle injuries (Anderson 2000). As such, single examples of rare normal variants/abnormal/controversial cases were retained while duplicate examples of these were removed. The final list of diagnoses is detailed in Table 1.

Table 1 Diagnoses included in exercise

Radiograph type	*N*
Normal	
Normal	131
Normal variant	15
Abnormal	
Rule out Salter–Harris I fibula	36
Salter–Harris II fibula	7
Salter–Harris III/IV/V fibula	0
Salter–Harris I tibia	1
Salter–Harris II tibia	9
Salter–Harris III tibia	9
Salter–Harris IV tibia	5
Salter–Harris V tibia	0
Combined tibia/fibula	1
Other pathology—osteochondritis dissecans	1
Total	234

Cases consisted of three images (AP, Lateral, Mortise views) and these were downloaded, along with the final staff pediatric radiology report, from the institutional picture archiving and communications system. Radiographs were saved in Joint Photographic Experts Group format and a brief clinical history was written for each case based on the information present on the imaging requisition. All cases were then categorized as *normal* or *abnormal* based on the information provided by the official radiology report. A *normal* AXR was defined as a radiograph without a visible bony fracture and/or lack of soft tissue swelling over open growth plates in the distal tibia/fibula. Ankle radiographs that contained variations on the normal anatomy but did not require treatment or further investigation of any kind (e.g., extra epiphysis) were included within the normal AXR group. Finally, an *abnormal* AXR was defined as a film with a visible bony fracture and/or soft tissue swelling over open growth plates in the distal tibia/fibula. *Abnormal* films were further subclassified by diagnosis and the location of the abnormality on the image. The Salter-Harris I fracture of the distal tibia/fibula was defined as soft tissue swelling maximal over the normal/displaced/widened open growth plates, and the absence of a visible bony fracture. If there were any uncertainties about the accuracy of the diagnosis on the original radiology report it was reviewed with an independent staff pediatric radiologist that specializes in musculoskeletal imaging.

Online software application for presentation of radiograph cases

A website was developed using HTML, PHP and Flash. Secure entry was ensured via a participant name and password given to each participant. The software tracked their progress through the cases, and recorded responses to a mySQL database. The information from this database was later entered into the SPSS® software analysis package.

At the start, each participant was given some general information which included assurance of confidentiality, the purpose of the exercise, and some information on how to use the system. They were not provided with any information about the proportion of normal to abnormal cases or types of pathology in advance of participation. Cases were

Fig. 4 Screen capture from after the learner submits their answer. The *dot* represents the learner's designation of the abnormality, just inside the "hotspot" representing the correct location

then presented in a random order unique to each participant. For each case, the participant was presented first with a screen listing the presenting complaint and clinical findings of the patient. Clicking the appropriate button took the participant to one of the three standard radiograph views of the ankle. The participant could access all three views as (s)he wished. No time limitation was imposed. When ready, the participant declared the case either "Normal" or "Abnormal" with modifiers suggesting how certain they were in the diagnosis. If the answer was that the radiograph is "Abnormal," the participant then marked the radiograph, using a yellow dot, to indicate where they thought the abnormality was located. They then committed to their answer by clicking a "Submit" button that lead to instantaneous feedback including a visual overlay indicating the region of abnormality (if any) and presentation of the entire official radiology report. An example of this screen is shown in the screen capture (Fig. 4). Once the participant has considered this information, they moved onto the next case.

Deliberate practice

The characteristics of this study consistent with deliberate practice as defined by Ericsson (Ericsson et al. 1993) include the following: (1) all participants were motivated due to high relevance of the task in their field, and/or the fact that they volunteered to participate; (2) ankle x-rays and their interpretation took into account subject pre-existing knowledge; (3) subjects received immediate feedback with knowledge of their performance; (4) the overall

task of ankle x-ray interpretation remained the same, and the specifics of each x-ray offered opportunities to improve.

Analyses

Each case completed by a participant was considered one item. Normal items were scored dichotomously depending on the match between the participant's response and the original radiology report. Abnormal items were scored correct if the participant had both classified it as abnormal and indicated the correct region of abnormality on at least one of the images of the case.

Descriptive statistics were used to report mean and standard deviation scores for continuous variables and proportions with respective 95% confidence intervals for categorical data. Discrimination (d') and criterion (λ) parameters were calculated using the equal variance Gaussian SDT model described in Wickens (Wickens 2002). For calculating the signal detection λ, there are several different forms reported in the literature (Wickens 2002). These include lambda, lambda-center (different referent d') and log-beta lambda (based on a likelihood ratio). The latter two have the advantage of being relatively independent of d'. In this paper, we chose to report raw lambda scores to provide greater transparency for the reader and better represent the sensitivity/specificity trade-off.

Our expectation was that the subjects would show improvement with successive practice, and that this could be captured by running estimate of the d' and λ parameters. With each additional exercise case, we automatically computed d' and λ, and then graphed them as a function of case. The resulting learning curves (Ericsson 2006) provide a formative analysis of learning with each case, and were plotted by graphing a cumulative calculation of the outcome variable against the sequence number. Since results are weighed down by the early cases where performance would not have been as good, these cumulative statistics underestimate the final performance level of each subject. However, our goal was not absolute assessment of performance; rather, we wanted to report, at a fine level of granularity, the relative changes in performance between groups as they learn.

To test overall statistical predictions, we also reported summative point estimates of d' and λ.

Summative statistics by groups were compared using ANOVA and post-hoc comparisons between groups were performed using the Tukey test. All analyses were carried out using SPSS 13.0, Stata 10.1 (College Station, TX) and Excel 2003 (Microsoft Corp, Bellevue, WA).

Results

Study participants

Forty-six participants with varying degrees of interpretative skill completed the study: 20 MED (6 University of Toronto, 12 Queen's University, 2 Columbia University), 6 RES (6 Morgan Stanley Children's), 12 FEL (7 Hospital for Sick Children, 5 Morgan Stanley Children's), 5 PEM (3 Hospital for Sick Children, 2 Morgan Stanley Children's), and 3 RAD (Morgan Stanley Children's).

Summative performance statistics

The discrimination ability (d') was higher for more experienced learners demonstrating that the latter group could better distinguish normal from abnormal radiographs (Table 2). At the end of the exercise, the bias towards reporting more films as normal was greatest in the high experience groups, while the low experience groups had a relative tendency to report films as abnormal.

The overall ANOVA was statistically significant for both d' and lambda. For d', pairwise comparisons of differences between the groups were statistically significant at $p < 0.05$ except for the comparisons between a) medical students and residents b) residents and PEM fellows and c) PEM attendings and radiologists. However, these non-significant post-hoc comparisons were all underpowered (Retrospective Power < 0.25).

Group level learning curves (Fig. 5)

Discrimination curves

In the low experience groups, d' improved with each case encountered. On average, there was a negative exponential pattern with rapid initial learning followed by slower incremental improvement. Improvement continued out to the final 234th case. For the high experience groups, d' improved rapidly and then levelled off by the 75–100th case and was stable thereafter with little, if any, further improvement.

Criterion curves

The high experience PEM and RAD groups, after the initial 50–75 cases, maintained a stable criterion, or balance between sensitivity and specificity. Their λ was fixed at about two-thirds of d'. However, for the lower experience trainees there appeared to be a developmental pattern. The medical students had a tendency to call films normal ($\lambda > d'$) using a very "strict" criterion. Lambda is high relative to d' when an individual's errors are predominantly due to abnormal images being classified as normal (i.e., high specificity and low sensitivity). This pattern persisted right out to the 234th repetition. The RES group had this strict criterion initially, but with increased exposure not only did their discrimination ability improve but their relative criterion shifted to more of the high experience pattern so

Table 2 Summative performance statistics

Variable	Medical students ($n = 20$)	Residents ($n = 6$)	PEM fellows ($n = 12$)	PEM staff ($n = 5$)	Radiology staff ($n = 3$)	p-Value*
Detection parameter dprime (95% CI)	0.68 (0.59, 0.77)	0.97 (0.89, 1.05)	1.13 (1.03, 1.23)	1.51 (1.39, 1.63)	1.76 (1.72, 1.80)	<0.0001
Criterion parameter lambda (95% CI)	0.45 (0.44, 0.47)	0.69 (0.67, 0.71)	0.87 (0.85, 0.88)	1.41 (1.38, 1.43)	1.65 (1.63, 1.67)	<0.0001

* Groups compared using ANOVA with 45 degrees of freedom and assuming $\alpha = 0.05$

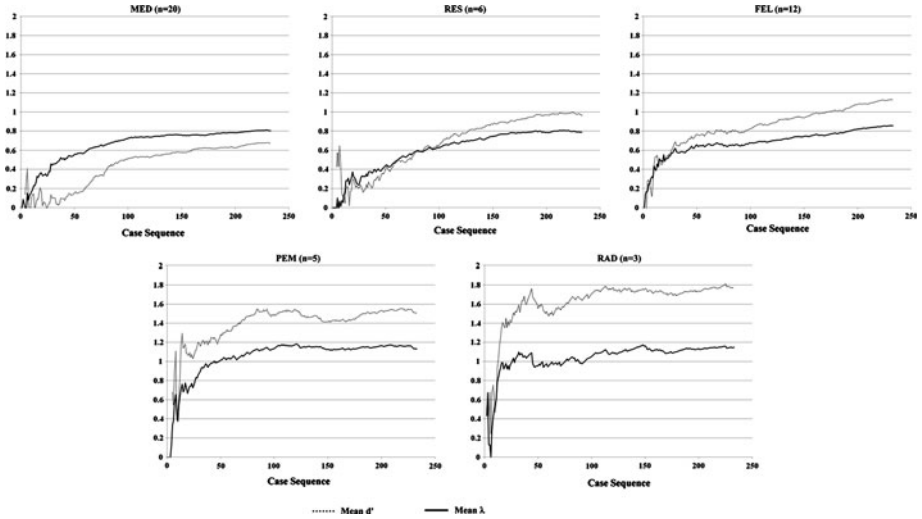


Fig. 5 Group level SDT learning curves for medical students (*MED*), residents (*RES*), pediatric emergency medicine fellows (*FEL*), PEM staff (*PEM*) and radiologists (*RAD*). Learning curve parameters d' and lambda are plotted on the *y*-axis against number of cases reviewed on the *x*-axis (maximum 234)

that $\lambda < d'$. The FEL group also showed an initially strict criterion that softened with case exposure so that the high experience pattern was well established by the last case.

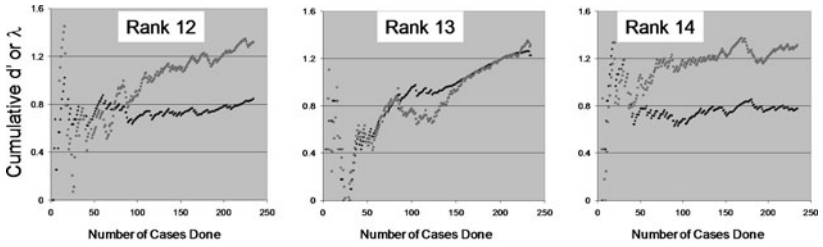
Individual level learning curves

For each individual participant, we plotted the SDT parameters against number of cases done. We then ordered these by the participant's final d' achieved. As expected, participants with low d', generally medical students, often showed the inverted pattern where $\lambda > d'$ as seen in the medical student group level curves (Fig. 5). However, in several instances adjacent cases (i.e., with similar final d') showed markedly different qualitative patterns with one showing the low experience pattern of $\lambda > d'$ while the other the high experience of $d' > \lambda$ pattern (Fig. 6).

SDT learning curves from three individuals each with the same final d' value of 1.3 are shown in Fig. 6. The medical student (case 13), has the same discrimination ability as the two fellows. However, (s)he uses a highly specific strategy (strict criterion), at the cost of misclassifying a greater proportion of patients with abnormalities (i.e., poorer sensitivity).

Discussion

Using learning methods emphasizing deliberate practice (Colvin 2008; Ericsson et al. 1993), this study examined changes in signal detection parameters with serial exposure to hundreds of ankle radiograph cases. Discrimination (d') improved with the number of cases reviewed for all groups, although the improvements were relatively small for the staff level practitioners. Increased ability to discriminate was concordant with level of participant seniority. We also determined how the participants set their internal criterion or bias when faced with a case in which they are uncertain. The low experience groups set their criterion



Parameter	Rank 12	Rank 13	Rank 14
Trainee Level	Fellow	Medical Student	Fellow
Cumulative d'	1.32	1.31	1.31
Cumulative Sensitivity	0.68	0.53	0.70
Cumulative Specificity	0.80	0.89	0.78

Fig. 6 Learning curves from three participants who ranked nearly identically in terms of their cumulative discrimination parameter (d') but different development of their criterion parameter (λ)

high relative to their discrimination ability which was in contrast to how the staff level physicians responded.

The discrimination pattern observed in our learning exercise was as one would expect. On average, discrimination improved with number of cases reviewed except for the radiology experts who had little to learn from this intervention. Construct validity of our exercise as a measure of interpretation ability is supported by the fact that groups with increasing levels of expertise showed concordant increased ability to discriminate. The medical student group was challenged the most throughout the exercise, demonstrating relatively poor differentiating abilities, even by the end of the experience. In addition, the PEM d' failed to achieve that of the RAD group even after our 234 item learning intervention. This study suggests that future research should be focused more directly on how this type of serial learning can be modified so that it can help all groups achieve a given level of competency.

Feature discrimination is not the only learning task in the skill of radiograph interpretation. Even experts have to learn to trade-off sensitivity with specificity when faced with ambiguous or borderline visual features on radiographs. A participant's approach to this task is summarized by the SDT criterion parameter. In this study, we found that those in the low experience group, especially the medical students, had a high criterion relative to their ability to discriminate. That this group had poor discrimination ability contributes to this. They would be more likely to miss an abnormality on the radiograph resulting in more false negatives and a higher criterion value. Conversely, as a person's ability to discriminate improves, there would be fewer false negatives and a relatively lower criterion value. Our findings support this since the latter was evident in the high experience group. However, poor discrimination is not the only factor as it does not explain the instances in our data where individuals of similar discrimination ability had markedly different criterion values. With our study design, we can only speculate as to why two individuals of similar discrimination ability would differ in how they balance sensitivity with specificity. It could be due to random variation between individuals or the differential influence of base-rate biases (McNicol 2004). Importantly, different perceptions of the cost

of false positives versus cost of false negatives needs to be considered. Finally, differences in motivation for the persistent searching required for detecting subtle abnormalities may also be a factor. While determining which factors are relevant will require further research, at this stage we can say that measuring a learner's criterion parameter allows us to detect those learners who over- or under-call pathology relative to their peers.

This research has limitations that warrant consideration. Not all skills lend themselves to this type of analysis. In this ankle radiograph example, the task is well codified and consistent from case to case, and knowledge gained from assessing one case is immediately applicable to the next one. The dichotomous nature of the problem, whether a fracture is present or not, allows for unambiguous feedback that may not be the case for other situations, for example chest radiographs taken to rule out pneumonia. We approached over 100 potential participants and only 46 agree to participate, which raises the concern of responder bias. Learners who participated may represent a different skill level than those who did not. There were small sample sizes in each group, and as such a few highly unmotivated or motivated learners in each group could significantly bias the results. Finally, small numbers of participants makes it difficult to perform subgroup analyses to identify the strengths/weaknesses of a given group.

In conclusion, signal detection metrics applied serially can provide valuable insight on changes associated with learning radiograph interpretation. Discrimination measured ability to differentiate normal from abnormal, and improved in all learner groups with the deliberate practice of a long series of radiographs. Predictably, this was higher in the more experienced groups. Criterion results provided information on tendencies to over or under call pathology. The low experience group was biased towards labelling radiographs as normal which missed pathology while the high experience groups set their criterion such that false negatives and false positives were minimized. Therefore, SDT parameters capture learner performance beyond the traditional "correct/incorrect" data, and this information may be incorporated to design more effective instructional strategies specific to a group or individual for a common and important diagnostic skill.

Acknowledgments We would like to thank the Royal College of Physicians and Surgeons of Canada for their grant support of this research.

References

Abdi, H. (2009). *Encylopedia of education*. New York: Elsevier.

Anderson, A. (2000). Injury—ankle. In I. G. Fleisher, S. Ludwig, F. Henretig, R. Ruddy, & B. Silverman (Eds.), *Textbook of pediatric emergency medicine* (pp. 321–329). Philadelphia: Lippincott Williams & Wilkins.

Boutis, K., et al. (2001). Sensitivity of a clinical examination to predict the need for radiography in children with ankle injuries: A prospective study. *The Lancet, 358*, 2118–2121.

Chung, S., et al. (2004). Skull radiograph interpretation of children less than age two: How good are pediatric emergency physicians? *Annals of Emergency Medicine, 43*, 717–722.

Clarkson, E. (2007). Estimation receiver operating characteristic curve and ideal observers for combined detection/estimation tasks. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 24*, B91–B98.

Colvin, G. (2008). *Talent is overrated: What really separates world-class performers from everybody else*. New York: Penguin Group.

Doubilet, P. M. (1988). Statistical techniques for medical decision making: Applications to diagnostic radiology. *AJR. American Journal of Roentgenology, 150*, 745–750.

Dowling, S., et al. (2005). Comparison views to diagnose elbow injuries in children: A survey of Canadian non-pediatric emergency physicians. *Canadian Journal of Emergency Medical Care, 7*, 237–240.

Ericsson, K. A. (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge: Cambridge University Press.
Ericsson, K. A., et al. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review—New York, 100*, 363–406.
Hatala, R. M., et al. (2003). Practice makes perfect: The critical role of mixed practice in the acquisition of ECG interpretation skills. *Advances in Health Sciences Education, 8*, 17–26.
McNicol, D. (2004). *A primer of signal detection theory*. New York: Routledge.
Metz, C. E. (1986). ROC methodology in radiologic imaging. *Investigative Radiology, 21*, 720–733.
Minnes, B. G., et al. (2005). Agreement in the interpretation of extremity radiographs of injured children and adolescents. *Academic Emergency Medicine, 2*, 826–830.
Nodine, C. F., et al. (1999). How experience and training influence mammography expertise. *Academic Radiology, 6*, 575–585.
Norman, G. R., et al. (1992). Expertise in visual diagnosis: A review of the literature. *Academic Medicine, 67*, S78–S83.
Obuchowski, N. A. (2003). Receiver operating curves and their use in radiology. *Radiology, 229*, 3–8.
Ramsay, C. R., et al. (2001). Statistical assessment of the learning curves of health technologies. *Health Technology Assessment, 5*, 1–79.
Reeder, B. M., et al. (2004). Referral patterns to a pediatric orthopedic clinic: Implications for education and practice. *Pediatrics, 113*, 714–719.
Ryan, L. M., et al. (2004). Recognition and management of pediatric fractures by pediatric residents. *Pediatrics, 114*, 1530–1533.
Swensson, R. G. (1996). Unified measurement of observer performance in detecting and localizing target objects on images. *Medical Physics, 23*, 1709–1725.
Taras, H. L. (1990). Ten years of graduates evaluate a pediatric residency program. *American Journal of Diseases of Children, 144*, 1102–1105.
Trainor, J. L., & Krug, S. E. (2000). The training of pediatric residents in the care of acutely ill and injured children. *Archives of Pediatrics and Adolescent Medicine, 154*, 1154–1159.
Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.